

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE	3. REPORT TYPE AND DATES COVERED	
4. TITLE AND SUBTITLE The Application of Artificial Neural Networks to Object Direction In Digital Images			5. FUNDING NUMBERS	
6. AUTHOR(S) Steady Warren Housholder				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AFIT Students Attending: University of Texas at Austin			8. PERFORMING ORGANIZATION REPORT NUMBER AFIT/CI/CIA 95-74	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) DEPARTMENT OF THE AIR FORCE AFIT/CI 2950 P STREET, BLDG 125 WRIGHT-PATTERSON AFB OH 45433-7765			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for Public Release IAW AFR 190- Distribution Unlimited BRIAN D. Gauthier, MSgt, USAF Chief Administration			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words)				
19950912 003				
14. SUBJECT TERMS			15. NUMBER OF PAGES 62	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT	

Abstract

THE APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO OBJECT DETECTION IN DIGITAL IMAGES

By

STEACY WARREN HOUSHOLDER, M.S.E.

Second Lieutenant, USAF

Supervising Professor: Dr. Joydeep Ghosh

University of Texas at Austin, 1995

This report surveys the topic of object detection in digital images. It outlines a standard method of analysis for the perception process. Several imaging and recognition issues are identified and explored through the human visual system as a model which is then used for constructing an artificial perception system. This leads to the study of self-organized artificial neural networks. The study includes the standard Kohonen feature map and several of its variations. The report will next focus on the application of these neural networks to object recognition tasks. The length of the work is 61 pages and includes research conducted in a wide variety of publications. Key primary and secondary resources are listed below (The original list is 34).

- [1] Stephen Grossberg. "Neural Pattern Discrimination." In Gail A. Carpenter and Stephen Grossberg, editors, *Pattern Recognition by Self-Organizing Neural Networks*, pp. 1-34. Cambridge: The MIT Press, 1991.
- [2] Ralph Linsker. "Self-organization in a Perceptual Network." In *IEEE Computer*, vol. 21, no. 3, pp. 105-117, 1988.
- [3] David Marr. *Vision; A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman and Company, 1982.
- [4] Erkki Oja. "Self-Organizing Maps and Computer Vision." In Harry Wechsler, editor, *Neural Networks for Perception; Volume 1; Human and Machine Perception*, pp. 368-385. New York: Academic Press, Inc, 1992.

Accession For	
NTIS CRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification _____	
By _____	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

Copyright

by

Steacy Warren Housholder

1995

**THE APPLICATION OF ARTIFICIAL NEURAL NETWORKS
TO OBJECT DETECTION IN DIGITAL IMAGES**

by

STEACY WARREN HOUSHOLDER, B.S.

REPORT

Presented to the Faculty of the Graduate School
of The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the degree of

MASTER OF SCIENCE IN ENGINEERING

THE UNIVERSITY OF TEXAS AT AUSTIN

AUGUST, 1995

**THE APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO
OBJECT DETECTION IN DIGITAL IMAGES**

Approved:

Supervisor: Joydeep Ghosh
Dr. J. Ghosh

B.F. Womack
Dr. B.F. Womack

To My Wife, Lisa Anne

Preface

I would like to take the opportunity to thank the faculty and staff of the Electrical and Computer Engineering Department of this university for all of their assistance and insight during the course of my studies. I would like to extend special thanks to my academic advisor, Dr. J. Ghosh, for his contributions to the research conducted in this report. Finally, I would like to thank my friends and family for all of their support, encouragement, and steadfast patience throughout my various studies.

Steacy Warren Housholder

The University of Texas at Austin

August 1995

Abstract

THE APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO OBJECT DETECTION IN DIGITAL IMAGES

By

STEACY WARREN HOUSHOLDER, M.S.E.

Supervising Professor: Dr. Joydeep Ghosh

This report surveys the topic of object detection in digital images. It outlines a standard method of analysis for the perception process. Several imaging and recognition issues are identified and explored through the human visual system as a model which is then used for constructing an artificial perception system. This leads to the study of self-organized artificial neural networks. The study includes the Kohonen feature map and several of its variations. The report will next focus on the application of these neural networks to object recognition tasks.

Table of Contents

Preface	v
Abstract	vi
Table of Contents	vii
List of Figures	x
1. Introduction	1
1.1 Purpose.....	1
1.2 Approach.....	7
1.3 Terminology.....	8
 2. Biological Foundations	 10
2.1 Chapter Outline.....	10
2.2 Human Visual System - Overview.....	11
2.3 General Imaging Considerations.....	12
2.3.1 Analysis Approach.....	13
2.3.2 Computational Theory.....	13
2.3.3 Data Representation.....	14
2.3.4 Invariance.....	19

2.4 Human Visual System - Hardware Implementation.....	21
2.4.1 Retina.....	23
2.4.2 Optic Nerve.....	25
2.4.3 Lateral Geniculate Nucleus.....	25
2.4.4 Visual Cortex.....	26
2.5 Discussion.....	28
 3. Kohonen Model	 30
3.1 Chapter Overview.....	30
3.2 General Properties.....	30
3.3 Learning.....	34
3.4 Discussion.....	39
 4. Network Applications	 41
4.1 Chapter Overview.....	41
4.2 Representational Form.....	42
4.3 Algorithm Structure.....	47
4.4 Dedicated Hardware.....	54

5. Conclusions	56
Bibliography	58
Vita	62

List of Figures

2.4 The Human Visual System.....	21
2.4.1 Center-Surround Receptive Fields.....	24
3.2 1 Dimensional Kohonen Feature Map.....	32
3.2 "Mexican Hat Function".....	33
4.3 Hierarchical 1 Dimensional Kohonen Feature Map.....	48

Chapter 1

Introduction

1.1 Purpose

The desire to create cognitive machines that can think, learn, and understand like the human brain motivates the research conducted in this report. Such machines should be capable of interacting with the surrounding environment based on various inputs and stored memories. They should be able to react to novel situations, draw intelligent conclusions, and save pertinent information for future use. Such machines should also be capable of communicating these processes to the outside world in some natural language. It is safe to state that the required qualities of any future rational machine are closely related to the current understanding of human cognition. This is by default since human cognition is the best and most advanced model currently available to researchers.

Leonard Uhr suggests in his book, *Pattern Recognition, Learning, and Thought; Computer-Programmed Models of Higher Mental Processes*, that the key to building machines that are truly aware is to understand fully the process of human learning [32]. He points out that in a child, all the tools of learning are

already present in the brain. All that is required for the child to become an interactive and understanding adult is the exposure to various external stimuli. Uhr proposes to begin with understanding the function of the brain and then fit the corresponding structure. This is a tremendous simplification over any approach that tries to examine a structure of the brain and fit its function. Most parts of the brain are involved in an unknown number of tasks at any one time that may be completely independent of each other. With the learning approach, researchers can simplify functions to their most basic state and work towards higher levels of understanding. The human model could be used to provide insight for this function and possibly for determination of any required hardware.

Uhr admits that even this approach of trying to understand the learning process may not be simple enough by itself [32]. In order for learning to take place, there must be something to learn about. In light of this, Uhr suggests that an even more suitable starting point in the attempt to build a cognitive machine is to focus on sensation and perception. Sensation and perception are the first necessary steps in the learning process which are responsible for taking relevant data from the outside world and placing it into a suitable format for higher interpretation. It is perception that poses the greatest current challenge to researchers. As a result, perception will be the focus of the current study. Uhr

adds that an advantage of studying perception is that many similar problems can be found in higher levels of cognition. These problems include ".....'concept formation', 'symbolization', and 'problem-solving'" [32].

In recent years there has been a tremendous amount of research conducted in the area of image processing. Historically, image processing has referred to the low level tasks of noise filtering, deblurring, geometrical transformations, histogram equalization, thresholding, and local edge detection [21]. These are tasks important for sensation. The great interest in image processing is due in part to the wide proliferation and acceptance of computers in both the commercial and private sectors. Image processing is useful in a wide variety of applications including photography, surveillance, satellite imagery, and medical imagery. Computer proliferation has itself led to the creation of a present need and motivation to produce machines with some basic abilities of both sensation and perception. The primary difference in these machines over conventional computers is their ability to interpret and categorize data. This need has surfaced primarily in industry with possible uses ranging from quality control in VLSI chip manufacturing and robotics to automated weather forecasting. Needs also exist in applications like target recognition and MRI tissue classification [25]. The result is that research focusing on perception in vision

provides benefits in the short term for industrial applications as well as in the long term for the future construction of cognitive machines.

Vision, as a particular domain, lends itself quite well to the problem of understanding perception. A still image contains a tremendous amount of information about the surrounding environment. The human visual system is dominant among all of the senses and is extremely acute in interpreting this information. Visual dominance is evident in light of the disproportionate part that the human visual system plays (with respect to the other four senses) in responding to changes in the environment. Think of how people respond when faced with a change in what they are hearing, feeling, tasting, or smelling. The first instinct is to locate the cause of the change with the individual's eyes. This report will focus on a specific type of perception, object detection and identification in images. Since the sense of sight is so acute in humans, researchers can actually see what results should be solicited from testing on a machine or algorithm. The impact of these results is much more direct than what a set of numerical results would be, and allows researchers to identify specific difficulties experienced by the machine or algorithm. Another advantage of vision analysis is that images are easy to obtain and easy to process. A wide variety of image types can be found to explore the detection of objects under varying circumstances.

Paralleling research in image processing during recent years is research in artificial neural networks. These systems have been inspired by biology in both structure and functionality. Research in this field is motivated by the acknowledgment that the human brain computes data in a completely different manner than a conventional digital computer [13]. Instead of a linear execution of instructions, a neural network structures data hierarchically, combines it in parallel nonlinearly, and after several combining steps reaches some final result. This distributed form of computation allows neural networks to be good at classification and generalization tasks. They are extremely valuable in finding relationships between two domains where the mathematical mapping is unknown. Erkki Oja writes that neural networks are particularly suitable for use in a visual perception system for two reasons [21]. First, this is primarily an intermediate level of computation consisting of segmentation, feature extraction, shape analysis, and texture analysis tasks. These tasks have computational requirements that neither over- nor under-utilize neural network assets. Oja also points out that the degree of parallelism ($O(n)$) for these tasks is both achievable and practical considering current technology for implementations using neural networks. A final reason for employing neural networks is that their structure and functionality are similar enough to actual biological systems

that they do provide greater insight into the way that perception is actually accomplished by the human visual system.

Report research will specifically explore the use of self-organized neural networks for application in visual perception systems. This subset of the neural network field most closely approximates the actual operation of many biological systems (not just limited to vision). Past research has shown that self-organizing feature maps have been useful in demonstrating the formation of internal mental representations and general organization of the human visual system [18]. The limitations of supervised neural networks prevent their practical use in perception. Oja writes that supervised networks often require extensive preprocessing in order to place data into a usable form [21]. In a modular system (like the human visual system), it is difficult to specify the individual module functions. Unsupervised networks avoid these basic limitations and are better suited for use in vision applications. They also allow researchers a means to perform most intermediate level imaging tasks without having to incorporate specific apriori information (although this can be highly beneficial).

1.2 Approach

This report will explore the various techniques currently employed in the application of self-organizing networks to the problem of object detection in digital images. Section 1.3 of this introduction chapter will precisely define a few of the more important terms and concepts that will be used throughout the report. It is the author's intention that this will help alleviate some of the ambiguity that is present in the field due to the current lack of terminology standards. Chapter 2 will seek to build a solid foundation in the current understanding of the biological structure and function of the human visual system. First, it will explore some of the aspects of general imaging. This will be specifically in relationship to the recognition of objects in an image. A general approach to the analysis of the perception process will be outlined. Next, the decision path from light sensation in the retina to object recognition by the brain will be traced. This chapter will provide valuable insight into the perception process. Due to the biological motivation of the self-organizing networks employed, the chapter will also provide insight into various construction and operation aspects required for perception. Chapter 3 describes the Kohonen feature map which serves as a representative model of many self-organizing neural networks. Several properties will be outlined. Chapter 4 will

then explore some current issues in object detection including several applications of self-organizing networks. Finally, chapter 5 will present some conclusions and parting thoughts.

1.3 Terminology

The purpose of this section is to lay a brief, yet concise foundation of standard terminology to be used throughout the remainder of the report. An *image* will be used to refer to a 2 dimensional representation of a 3 dimensional environment. An *object* is a collection of connected edges and textures in the environment referred to as a single entity [33]. An image captures the 2 dimensional representations of a set of objects. The first step in identifying the presence of a particular object in an image is the sensation phase. *Sensation* refers to the simple one to one transformation of data from one representation to another that is more useful. In the human visual system, for instance, this is the transformation of light in the environment to electrical impulses in the retina. Sensation is relatively easy to replicate with conventional computers. *Segmentation* refers to the process of separating potential objects from the background of an image [14]. *Perception* is the task of selecting, organizing, generalizing, and classifying significant data. This is a many to one mapping

which reduces the data set to a smaller more information-rich representation. The goal is to capture some invariant aspect of the surrounding environment [34]. This definition of perception is well accepted and can be found in [32], [19], and [34]. The term recognition is the combined process of sensation and perception. Two final terms need to be defined. A distinction must be made between pattern recognition and object recognition. *Pattern recognition* is the sensation and perception of a 2 dimensional figure in the environment. *Object recognition* is a similar function that is performed on 3 dimensional entity. The latter function is the more difficult to perform due to the increased number of parameters.

Chapter 2

Biological Foundations

2.1 Chapter Outline

The objective of this chapter is to lay a solid foundation for the biological understanding of how the human visual system performs the tasks of sensation and perception. The chapter will focus on the task of object recognition. Roger Watt describes the overall process of vision by the human visual system in his book, *Understanding Vision* [33]. He writes, "Vision is the extraction and analysis of information from an optical image in preparation for, and execution of, behavior within the scene" [33]. This definition is a specific application of the perception term defined in Chapter 1. The second chapter will begin by giving an overview of the way that the human visual system operates. Research will next emphasize the human visual system's ability to address key imaging issues that apply directly to object detection. An approach for analyzing the perception process will be outlined here for use throughout the report. The chapter will then trace the sensation and perception process as it permeates through the human visual system.

2.2 Human Visual System - Overview

The human visual system is a highly modular group of organs capable of acquiring and interpreting optical data from the environment. The system processes data in a top down, constructive fashion [34]. It operates like a selective filter choosing only relevant data to be passed on to subsequent processing levels. The human visual system (HVS) performs its processing tasks at several different image scales in parallel. This is evident from the way that a person can see edges, windows and doors, and buildings all at once in an image. Biologists agree that the HVS performs most of its tasks in parallel and are highly skeptical that any iteration takes place [20]. The first stage in the process is a set of filters that perform some of the basic low level visual tasks like deblurring, local edge detection, thresholding, and some primitive forms of noise filtering [21]. This is followed by a set of feature detectors that perform more intermediate levels of image processing. Feature detectors allow for a dimensional reduction in the data set while maintaining the same information content [13]. They are used in standard image processing to emphasize important attributes of input data while at the same time reducing bandwidth and storage requirements. These attributes are then used for other tasks like

shape analysis, texture analysis, segmentation, and finally object recognition. The final task of object recognition is performed using a combination of important features and what is referred to as image context. The human visual system is able to place a great amount of reliance on context due to the redundancy inherent in most of its recognition chores (written letters in words, objects present in certain environments, and spoken languages).

2.3 General Imaging Considerations

This next section will address a variety of basic imaging considerations important in object recognition. The section will also explain some of the conduct of the human visual system in relation to these considerations. The first step in the analysis of any process is to outline a strategy for its thorough investigation.

2.3.1 Analysis Approach

David Marr in his book, *Vision*, furnishes one such strategy [20]. He explains how the problem of object detection can be analyzed in three different modes. First, the process must be examined in light of its computational theory. This refers to defining the goal of the process and outlining a general flow of its operation. Second, the representation and algorithm of the process must be assessed. The representation of the input and output data must be specified as well as the algorithm. Finally, the hardware that is necessary to implement the process must be specified. These three items must be addressed in order to fully understand how any system successfully performs the task of perception. They are the same items that researchers must render precise in order to build a machine capable of perception.

2.3.2 Computational Theory

Chapter 1 gave a general description of the goal of the perception process. That goal is to identify specific objects in an image. Perception in the human visual system is a directed and active process [34]. Directed refers to the

observers ability to choose the information to be processed. Active is characteristic of exploration which is essential for the observer to be selective in information acquisition. Directed and active perception builds upon low-level invariant object representations and mappings. In addressing the second issue for this first mode of investigation, the general logic of the process is outlined above in Section 2.2.

2.3.3 Data Representation

The next task is to address the representation of input and output data as specified by the second mode of investigation. The issue of the actual process algorithm employed by the HVS to transform input data into output data will be dealt with in section 2.4 along with the hardware implementation. Data representation is one of the most crucial issues in perception. A particular depiction of data allows for the explicit representation of certain important features which are used to delineate between different objects. There is a difference between explicit representations and implicit representations. An explicit representations is a direct symbol of a feature or object. An implicit representation is an indirect symbol with the same information content but it must be further processed in order to be made explicit [8]. Data representations

define the level of complexity that must exist in a mapping relationship of input to output space. This will all vary depending on the particular object and environment. It is important to realize that this mapping relationship is a many to one mapping. There are usually multiple feature sets that defining an object. The human brain even allows for multiple mappings from different senses.

David Marr outlines three aspects that describe a representation scheme [20]. The first is the coordinate system that is utilized. A system that specifies the positions of objects in an image relative to the viewer utilizes a viewer-centered coordinate system. The alternative is to specify locations relative to a viewed object. This is referred to as an object-centered coordinate system. A viewer-centered coordinate system is easier to employ, but requires that same objects viewed from different angles and translations be classified as different objects. This results in a tremendous amount of storage overhead even for a small number of objects. A viewer-centered coordinate system is analogous to taking a picture of a scene and then trying to segment and identify the objects that are present. The object-centered approach avoids some of these memory requirements and makes object perception computationally easier. This system is comparable to taking a picture of a scene and generating an independent coordinate system for each possible object. Each can then be rotated and scaled to some standard format. Actual object identification can, as a result, always be

performed by comparison with standard templates. The weakness of the approach is that potential objects must be segmented from the image in preceding steps. Specification of a coordinate system only determines how the perception workload is to be divided up among processing levels.

The human visual system exploits a viewer-centered approach [8]. Michael Seibert and Allen Waxman report that there is significant biological support for this [29]. They reference studies in [22], [23], and [24] by Perrett and others. These studies show that specific cells in the brain of a macaque monkey were found to be active for both two dimensional and three dimensional representations of a particular face or head seen from a single perspective. Other cells were found to be active for the same object viewed from a different side or angle.

The second aspect of an image representation scheme includes primitives [20]. Primitives are the basic elements of shape information used in a process. There are two basic types, surface-based (for two dimensional applications) and volumetric-based (for three dimensional applications). A simple primitive may describe the location and size of small pieces of an image surface. This may be in the form of some basic edge or contour unit. A more complex primitive may describe orientation and depth information. The cost of a more complicated representation is the amount of work that will be required by preceding levels of

processing. Biologists are still unsure as to the exact manner of information representation employed by the HVS for inputs to the perception process. Information that is probably represented includes occluding contours and cues, surface contours and cues, surface orientation, surface texture, and various shading attributes. There is some debate as to whether or not the HVS utilizes optical flow information. The requirement that an effective primitive must satisfy is that it be capable of fully representing all relevant features necessary in the process it is being used in. These features must result in representational separability which is the central issue involved in recognition tasks [34].

A common primitive of spatial information is a contour. Biologists do agree that the HVS uses some form of this representation. David Marr identifies three different types of contours [20]. The first is an occluding contour which defines a discontinuity in depth. This contour can be thought of as the silhouette of an object. When the HVS is presented only silhouettes of an object, it incorporates apriori information in order to make an educated guess about the object's identity. A second type of contour follows discontinuities in surface orientation. The HVS has the tendency to assume that these interior contours are convex. The final type of contour is a surface contour. This is a contour that lies physically on a surface like a marking or shadow line. The HVS uses these basic contours to calculate other surface parameters like texture and shading.

Bart Romeny and Luc Florack point out that there is neurophysiological and psychophysical evidence that the HVS can even determine spatial derivatives up to the fourth order to determine some parameters [27].

The final aspect of an image representation scheme is the organization that it imposes on its primitives. This can be used to weight some primitives more than others. The 2 1/2 dimensional representation is one example of a representation scheme [20]. Its primitives are distances to surfaces, surface orientations, occluding contours, and contours along discontinuities in surface orientation. The scheme is more than a two dimensional representation since it contains information about depth [8]. It is less than a three dimensional representation since this information is not explicit (depth information is contained in gradient form). The organization of this scheme allows all elements to be weighted as equally important. The HVS utilizes a three dimensional representation in order to conduct the task of perception. Researchers propose that this representation may even use inputs from some form of the 2 1/2 dimensional scheme [8]. The inputs may also stem from groups of two dimensional representations taken from various viewpoints [9]. The organization of the 3 dimensional representation is based on modules. These modules consist of groupings of similar primitive types. Scientists postulate

that a person is not aware of any representation type [8]. A person is only aware of results.

2.3.4 Invariance

A final key issue in perception is invariance. As previously stated, the goal of perception is to capture some invariant aspect of the surrounding environment [34]. What makes this process so complicated is that input data is inherently variable. One object may take on an infinite number of parametric representations due to possible translations and rotations in a scene. This does not even include possible variations due to noise or distorted camera (human eye) imaging.

The human visual system has the dual task of identifying objects in spite of variation and interpreting the type (and sometimes even the cause) of this variation. Variation contains a certain amount of valuable information of its own. In any working perception system, there must be certain attainable and invariant features that can be isolated for the recognition of an object. Terry Caelli, Mario Ferraro, and Erhardt Barth explain that invariance can be satisfied to different degrees [3]. Strong invariance is obtained when a representation's features uniquely and completely define the patterns or objects in an image.

Weak invariance includes representations that may only be able to broadly categorize objects based on the features represented. The HVS employs a combination of strong and weak representations which forces researchers to cross over into the realm psychology. Weak representations may be used to identify less familiar objects or objects that are missing some feature information (due to noise, vantage point, etc.). The human visual system's robust nature is evident in its ability to fill in where information is lacking with apriori knowledge of a particular scene. This is used to make an educated guess about the identity of a highly distorted object.

2.4 Human Visual System - Specific Implementation

This section will explore the actual hardware implementation and specific perception algorithm of the human visual system (HVS). Figure 2.1 below illustrates the generalized path that will be traced.

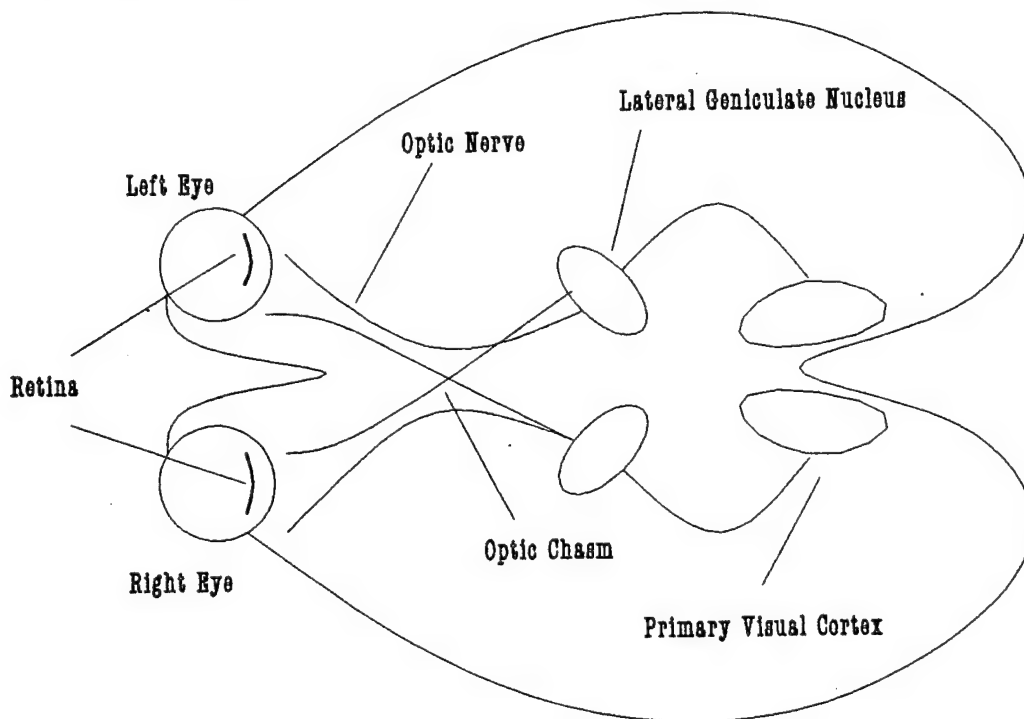


Figure 2.1 The Human Visual System

The first step in the perception process is the sensation of light in the environment. The sensitivity of the human eye to light stimulation is greatest about an area in the center of the visual field referred to as the fovea [15]. The HVS operates by directing the fovea at important points in the environment.

This is an active process of scanning the visual field for important information. The process of eye movement described above is referred to as gaze control [28]. Gaze control allows the information gathering resources of the eye to focus on one particular object or area of the visual field. The recognition of objects is accomplished using information gathered by the fovea. Periphery information is still gathered, but at a significantly lower resolution. This information is used for alerting, guidance, and gross feature identification. Gaze control is directed in three different ways [28]. It can be driven by a directed task, a voluntary fixation, or an involuntary reflex. It is thought that the parietal lobe may be responsible for moving the fovea about a scene. This control is a form of preprocessing for the HVS. It allows the viewer-centered HVS to exploit some of the more beneficial characteristics of an object-centered coordinate system. Objects that the HVS intends to identify are placed at the center of the visual field which is the origin of the viewer-centered coordinate system.

2.4.1 Retina

Once the attention of the fovea has been fixated on a particular portion of the visual field, it is the responsibility of the lens to focus the image onto the retina in the back of the eye. Photoreceptors in the retina absorb initial light photons. These receptors are referred to as the cones and rods. There are 6 million color sensitive cones inside the fovea [30]. 120 million rods, for peripheral and night vision, reside predominately outside the fovea. These cells convert light radiation into electrical impulses. Higher light intensities produce faster firing rates in these cells. Firing rates are the measure of intensity for any stimulation in the HVS. The impulses produced travel through a layer of bipolar cells, horizontal cells, and amacrine cells which serve in some early image processing tasks. These cells transmit their outputs to a layer of ganglion cells which are the cells responsible for data transmission along the optic nerve. These first two cell layers responsible for performing a considerable dimensional reduction on the inputs to the retina. This is evident in the fact that there are over 120 million photoreceptors and only 1 million ganglion cells to transmit information to the next stage of visual processing [30].

The formation of center-surround receptive fields are believed to be the major cause of this tremendous reduction. These areas contain information

about groups of photoreceptor cells in a single format. The fields were first described by Stephen W. Kuffler at the John Hopkins University School of Medicine in 1953 [15]. Center-surround receptive fields are formed by combinations of bipolar, horizontal, and amacrine cells. Their output is a single ganglion cell. The activation of the field must meet two conditions which are illustrated below in Figure 2.2.

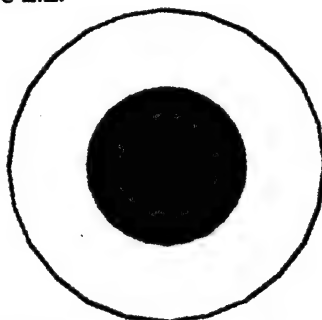


Figure 2.2 Center-Surround Receptive Fields

The center of the field must be illuminated while the surrounding area must not. Some fields are completely opposite being activated when the surrounding area is illuminated and the center is not. These receptive fields represent the first steps in the organization of data for perception tasks.

2.4.2 Optic Nerve

The axons of the ganglion cells in the retina form the optic nerve. One of the reasons that there is such a data reduction in the retina is that there is a bandwidth limitation imposed by the size of the optic nerve [30]. Routing of information in this nerve preserves the content of the left and right visual fields. These fields are kept separated by the retinas in both eyes. The optic nerve of each eye routes information regarding the right half of the visual field to the right hemisphere of the brain. The left half of the visual field is correspondingly routed to the left hemisphere of the brain. Portions of this nerve cross at what is called the optic chiasm.

2.4.3 Lateral Geniculate Nucleus

The next step in the perception process of the HVS is the optic nerve's connection to the lateral geniculate nucleus (LGN) [30]. The LGN is actually a part of the Thalamus which is responsible for directing information to appropriate areas in the brain [12]. The LGN further reduces the data set with six layers of center-surround fields. These fields are more receptive to color stimulus [30]. They discern between specific color combinations. Most of the

learning occurring in the HVS is thought to take place between the excitatory synapses of the LGN and the primary visual cortex [7].

2.4.4 Visual Cortex

The LGN leads directly through the parietal and temporal lobes to the primary visual cortex. A feedback path exists back to the LGN [15]. Its specific function is still unknown. The primary visual cortex is an extremely complex portion of the brain that is organized into six identifiable sheets or layers occupying an area of some thirty square centimeters [27]. It is a highly nonlinear network containing over 200 million cells. The primary visual cortex represents a direct mapping of features from the outside world. There are two primary transformations that are accomplished. The first is a transformation of information from a center-surround format to a line segment format. The second transformation includes a combination of information from both eyes [15].

The primary visual cortex is composed of four basic cell types that aid in both of its basic responsibilities. These are simple cells, complex cells, hypercomplex cells, and end-stopped cells [27]. Simple cells are orientation-selective. These cells are responsive to certain lines and edges present in specific

locations in an image. Complex cells are also orientation-selective. They are, however, responsive to lines and edges in any area in the receptive field oriented in the proper direction. These cells begin the process of combining information from both eyes. Hyper-complex cells consist of various combinations of complex cells. End-stopped cells compute boundaries in the same manner as zero-crossings. All of these cell types are interconnected on each of the cortex's layers. According to Dale's Principle each cell has either all excitatory or all inhibitory connections [11]. There is even evidence that lateral inhibitions are also present within the layers [7]. Combinations of these cells are capable of determining complicated features and even feature changes in the form of derivatives. This latter computation is valuable for determining texture boundaries and interior contours.

Processing in the primary visual cortex begins with simple orientation detection and proceeds through various levels of increasingly complex feature detection. This is accomplished through a hierarchy of levels composed of the basic cells described above. These cells self-organized into groups with similar characteristics and complexities, receptive field positions, orientations, and ocular dominance [15]. Information in each of the six layers is organized into vertical orientation columns [7] [16]. Locality is important in that groupings in close proximity are found to be iterated versions of each other. The use of

column organization is thought to be an attempt to represent a three dimensional environment on a series of two dimensional structures [15].

2.5 Discussion

The path of information that leads out of the primary visual cortex remains a mystery. How the brain uses orientation and feature information from the primary cortex is also unknown. All that can be clearly stated is that the cortex is not the final stage in the perception process [15]. There is research currently being conducted on higher cortical visual areas which are thought to receive projections from the primary cortex. In any case, the human brain is somehow able to construct more complicated features from information generated in the cortex. Apriori or context information is incorporated at some point in the process and aids in the final identification of objects. For a listing of actual performance qualities of the human visual system see [10] or [30].

This chapter shows that the human visual system is a highly complex system. It presents an overview of what is currently understood about how this system works in performing the tasks of sensation and perception. The HVS has the quality of being extremely modular and robust. An analysis approach for understanding aspects of this system and the perception problem is

presented. The chapter also shows how the human visual system addresses key imaging issues in relation to object detection. Finally, the current understanding of the perception problem is traced through the biological system. This chapter serves as a foundation for the understanding of perception as a process. This understanding is important for artificial systems as well as biological ones.

Chapter 3

Kohonen Model

3.1 Chapter Overview

The objective of this chapter is to introduce the topic of unsupervised neural networks for vision applications. These self-organizing networks are motivated by the biological considerations outlined in chapter 2. The reader can reference [13], [1], [26], or [14] for a good foundation in neural networks and basic self-organization. This chapter will highlight the Kohonen feature map which serves as a representative model of many self-organizing networks [1]. The chapter will also explore several optimization principles involved in self-organization.

3.2 General Properties

The guiding principle of self-organization is that global order can arise from local interactions [13]. Input patterns are presented to the network producing certain activity. The connecting weights between neurons in the

network are then modified in response to this activity. Weight modification or learning is based on a general form prescribed by Donald Hebb in 1949 [5]. Hebbian learning is a basic rule that is imposed on neurons in close proximity to each other. If one cell lies next to another and contributes to that cell's firing in a repeated and consistent manner, then the weight between the two cells is increased. The amount of increase in a weight is directly proportional to the degree of covariance that exists among all cells that affect a single cell's firing [13]. A form of Hebbian learning is believed to take place in the human visual system between the lateral geniculate nucleus and the primary visual cortex [7].

Kohonen networks apply this type of learning to an artificial neural network. The result is an unsupervised network in which there are no test patterns that can be used to find a mapping relationship between an input space and a specified output space. In fact there is no specification made of the output space. An unsupervised neural network produces a topographical representation of an input space. This mapping is an implicit and often reduced representation based on salient features [14]. These features allow the network to retain the same information content in representation as is in the input space. Similar features in the input map to adjacent neurons in the output layer. The network self-organizes into groups according to various features. This neighborhood or grouping characteristic also gives unsupervised networks

some error tolerance [21]. Unsupervised networks have been observed to self-organize into high degrees of structure in representing an input domain. Additionally, Kohonen writes that network resources are employed in an optimally efficient manner [16]. The strength of unsupervised neural networks is their ability to discriminate between various key aspects of the input space. This is essential for pattern and object recognition.

The Kohonen feature map usually consists of a one or two dimensional arrangement of interconnected neurons in a single output layer. An illustration of a one dimensional Kohonen map is shown below in Figure 3.1.

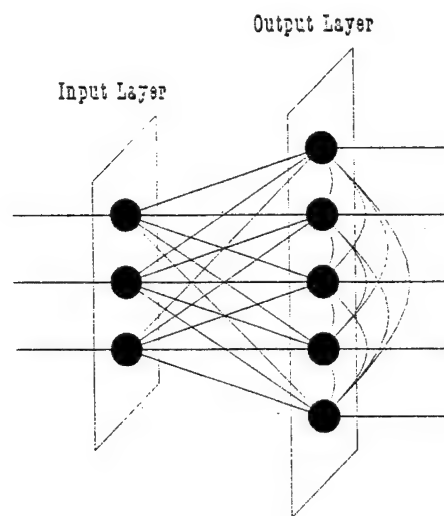


Figure 3.1 A 1 Dimensional Kohonen Feature Map

The size of this layer is chosen so as to minimize the amount of training time required while retaining adequate space to allow for the representation of all

salient features [1]. The input layer can be of any dimension and size. It is usually fully connected to the output. Apriori information about the input space may be used to alter this connectivity. This can result in significantly faster and more accurate self-organization of the network.

The interaction of neurons in the output lattice responding to an input stimulation can best be described by the "Mexican hat function" [1]. This activation function specifies three regions of lateral influence with respect to any single neuron. The interaction is illustrated below in Figure 3.2.

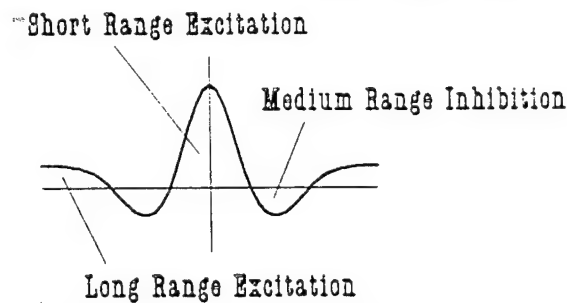


Figure 3.2 The "Mexican Hat Function"

These regions are all local to the neuron such that neurons on opposite sides of the lattice are independent and have no influence on each other. Neurons in the local region that are immediately next to a particular neuron provide a strong lateral excitation to that neuron. Neurons that are at medium distances cause a small amount of inhibition to the neuron. Neurons that are farthest away in the local area cause small excitations. This type of lateral feedback is similar to that

found in the center-surround receptive fields of the human visual system. Equation 3.1 below illustrates the response of a single neuron [13].

$$y = \phi \left(\sum_{i=1}^R w_i x_i + \sum_{k=-K}^K C_k y_k \right) \quad \text{Equation 3.1}$$

Let the response of an arbitrary neuron be y . The signals from the input layer will be denoted by x_i while R is the number of inputs. The synaptic weights of these inputs are represented by w_i . The lateral interconnections described above are denoted by C_k . The variable K includes the radius of neurons in the local region. The function ϕ denotes any nonlinear function that limits the output (y) and prevents it from assuming negative values. This lateral interaction of neurons in the output layer is present both during and after learning.

3.3 Learning

Self-organization is accomplished between the input and output layers by modification of connecting weights. The objective is to apply a local modification rule that will result in the emphasis of important features. Simon Haykin writes in *Neural Networks; A Comprehensive Foundation* that in order to achieve self-organization and not simply stabilization of the network these weight modifications (w_i) must tend to self-amplify [13]. This is consistent with

Dale's Principle which states that all weights of a neuron must either be excitatory or inhibitory [11]. Note that there are some artificial networks that do utilize inhibitory connections consistent with this principle. In order to prevent unbounded growth of a network's weights, neurons are forced to compete for a limited amount of resources. In other words only a few neurons are allowed to fire at any one time. This is constrained by only modifying the weights of the most excited neurons. In the Kohonen model weight modification is further constrained so that only the neuron with the greatest excitation has its weights increased. This is all referred to as competitive learning. Since the neurons are competing in a winner-take-all manner, no single neuron is able to fire without the assistance of the neurons around it. This means that there is a tendency for weight modifications to cooperate resulting in areas of strong weights for particular features of the input space.

These principles are applied to the learning process as follows. First a random set of weights is assigned to the synaptic connections between the input and output layers (w_i). Small magnitudes tend to prevent any initial bias allowing for increased discrimination between features in the input space. The next step is to introduce a sample vector (x) from the input space to the network. The vector should be chosen in direct relation to its probability of occurrence in the input space. The third step is to choose the winning neuron in the output

layer. This can be done by using an equation of the form shown below in Equation 3.2 [13].

$$y_{win} = \arg_j \min \|x - w_j\|, \quad j = 1, 2, \dots, N \quad \text{Equation 3.2}$$

N denotes the number of neurons in the output lattice. This equation chooses the minimum Euclidean distance that exists between the input vector and each of the weight vectors. The fourth step in the learning process is to modify the appropriate weights in the network. This is accomplished using Equation 3.3 [13].

$$w_j(n+1) = \begin{cases} w_j(n) + \eta(n)[x(n) - w_j(n)], & j \in \Lambda(n) \\ w_j(n), & \text{Else} \end{cases} \quad \text{Equation 3.3}$$

The current iteration is denoted by n . Robert Hecht-Nielsen describes the Kohonen learning process as a movement of the weight vector for the winning neuron from its original position towards the input vector (x) [14]. This is accomplished for all weights in the neighborhood (Λ) of the winning neuron (y_{win}). The amount of this movement or weight modification is controlled by the parameter η . Both the learning parameter and the neighborhood size are varied with time in order to aid in learning and map formation. The learning rate parameter should be decreased from unity to .1 over time [13]. This allows for broad adjustments as the map is initially forming. Once the general map has

been constructed, decreasing the parameter allows for a more detailed "fine tuning" of the network. The neighborhood parameter should also be relatively large at the onset of training to allow for an emphasis on strong lateral feedback. This causes the weight vectors to be highly correlated which results in an initial smoothness in the map surface. As the neighborhood size is decreased over time, the emphasis shifts to inhibiting or negative feedback. This allows for the uncorrelated growth of weight vectors which results in rough areas and greater detail in the map.

This entire learning process of weight modification is repeated until an adequate representation of the input space is made. The amount of time that is required in order to fully train a network is referred to as the saturation time. At saturation, the state of the network remains constant even with continued presentation of training patterns. In this state, the network has constructed its optimal representation of the input space. Saturation time can be decreased with the use of the "n-way" algorithm [1]. This algorithm makes use of the clustering features of the Kohonen model. This is simply a divide and conquer technique employed using a number of response windows. A "representative" neuron is chose from each of the n windows. The winning neuron among these "representatives" has its window examined in the next step. This winning

window is then divided into n smaller windows. This process continues until the size of the windows has been reduced to one neuron, the final winner.

Since the dimensionality of the lattice in the output layer is usually less than that of the input layer, several features are employed to define the complete representation. The final set of weight vectors (w) is most dense in areas where the input vectors (x) are most numerous [14]. As a result, it is vital to include a representative sample of the actual input space in the set of inputs used for learning. The lack of conformity to the probability density function of the input space can distort the mapping relationship of the Kohonen model [14]. This can be compounded by another source of distortion in feature maps. The distortion is a result of the map's inherent tendency to over represent areas of low density inputs and under represent areas of high density.

3.4 Discussion

This chapter has outlined in brief the Kohonen model of unsupervised neural networks. The model represents a first step in the construction of perception systems patterned after human biology in both structure and learning algorithm. The general performance of Kohonen-type networks is surprisingly similar to that of the human visual system. Hayder Ali Alkasimi in his thesis, *Explorations in Cognitive Processing and Visual Neural Recognition*, found that over a number of training presentations, the number of weight vectors that were within a certain radius of the input vectors followed an 'S' curve. This is to be expected since at the onset of training few of the weight vectors are close to the input vectors. Over time the number of close weight vectors increases. The rate of this change initially increases and then with additional training decreases to zero as the number of close weight vectors reaches a constant plateau. The shape of the entire function resembles that of a hyperbolic tangent. Another similarity can be seen in the final output state of a Kohonen network. The network is a topographical mapping of the input space. This is the same result that has been observed in higher layers of the visual cortex. One important difference between digital and biological systems is the

manner in which forgetting is carried out. A disadvantage with all current neural network implementations is that when storage capacity is reached, new inputs erase old information. In the human visual system, new information is continuously combined with old information thus preserving the total information content of the system. Current research into Kohonen models has been extended to include the development of hierarchically structured networks similar to those found in the human visual system [1]. Unsupervised neural networks may not provide the final solution to the perception problem. They may only be part of the solution. Erkki Oja points out in his article, "Self-Organizing Maps and Computer Vision", that these networks, at the very least, do provide valuable insights that may lend clues to other more efficient approaches [21].

Chapter 4

Network Applications

4.1 Chapter Overview

This chapter will build upon the previous two and investigate current applications of artificial neural networks to the perception problem. Although the objective of this report is to address object detection, many techniques can be borrowed from pattern recognition applications. Applications discussed in this chapter will, therefore, include both types. These applications will be examined in light of the strategy outlined in chapter 2. Most of the current research focuses on the second and third aspects of this strategy. These aspects include the representation of data, the general algorithm structure, and the construction of dedicated hardware. Section 4.2 will describe some current data representations that are being proposed. Several algorithm organizations will be considered in section 4.3. Finally, section 4.4 will examine in brief a dedicated sensory device designed to mimic various characteristics of the human eye.

4.2 Representational Form

The representation of data is critical in enabling any system to discriminate between objects in a scene. Of the various aspects of a representation scheme discussed in chapter 2, only the primitive construction and organization provide for practical avenues of future research. The goal of a good primitive is to accent important differences in input data. In designing a primitive, a researcher uses apriori knowledge to specify which feature types in a scene contain relevant information. This can both simplify a task and limit the resulting system's applicability to other tasks. For example, occluding contours may be useful as primitives in some 2 dimensional recognition tasks. Applied in for 3 dimensional recognition, these primitives could not fully specify information about an object. They would only be able to specify its shadow. All other information would be lost. A primitive simply imposes some basic structure on the information that exists in a scene. Information is usually first collected as raw data in a pixel intensity format like a gray level. Researchers must then perform some degree of preprocessing to convert this into structured data. Currently, there is no ideal representation scheme.

Janusz Starzyk and Sinkuo Chai introduce a primitive construct based on angle information derived from the spatial relationships of pixel intensities [31]. This is called a vector contour representation (VCR). A key issue in any recognition system is invariance. The authors acknowledge this importance and design VCR to capture scale, rotation, and translation invariant qualities of binary images. First, specification is made of a set of standard angle templates (0° , 90° , -90° , etc.). A template denoting segment or pattern end points is also included. These templates are used to encode the angular relationships of neighboring pixels in an image. Encoding begins at any location in the image and proceeds until all the angular relationships are defined by a set of templates. This encoding is the same as chain coding used in image processing to reduce the storage requirements of contour based patterns. Chain coding effectively reduces a two dimensional representation to a one dimensional occluding contour vector of a pattern or object. This implicit representation is a local definition of pixel relationships and as such contains a significant amount of noise. Noise can be removed using an averaging filter to perform some degree of smoothing. The authors employ a windowing function that averages the angles of neighborhoods of four templates. The authors next integrate angle

values in order to calculate accumulative angular variation. This is followed by a normalization of both the magnitude and length of the vector.

Normalization here results in scale and translation invariance. The final result is a one dimensional primitive of a two dimensional pattern. The authors next take the FFT of the vector which adds rotational invariance to the representation.

The authors apply VCR in both supervised and unsupervised networks for character recognition only [31]. Although, the authors describe their representation for use in object recognition, there is no evidence that three dimensional applications could be possible. Since there is only one primitive type, there is no requirement for any additional primitive organization. The performance of this representation for character recognition is reported to vary between 94% and 100%. This is performed on images containing single patterns only. It is possible that blob counting could be employed to deal with multiple characters in an image. This primitive definition is limited, however, to two dimensional pattern recognition due to its sole reliance on occluding contours. Occluding contours are not invariant in three dimensions and can in fact be highly misleading in what they describe. An example of this is how the shadow of a six sided cube can appear as a hexagon or a square depending on vantage

point. The representation would also encounter difficulties with gray scale or color images where thresholding results in inaccurate descriptions of elements in an image. Finally, the representation is extremely sensitive to the types of patterns being recognized. Characters are relatively simple patterns to represent since they contain few outlying segments and generally require only one contour to describe. Chain coding often experiences problems with patterns containing multiple open end and/or crossing contours. This representation is, therefore, not very robust.

Marijke Augusteijn and Tammy Skufca present another implicit primitive representation that incorporates texture information [2]. Textures are formed by gray levels appearing in an image in some periodic form. This information is applied to an object recognition task of finding human faces in a scene. The authors gather pictures of human faces taken from various angles and distances. Texture features from these pictures are derived from various gray level statistics. The authors first use co-occurrence matrices to measure the frequency of specific gray levels in four orientation directions (horizontal, vertical, and both diagonal directions). These matrices allow for rotation and translation invariance in that they simply represent the presence of textures. Eight types of second order statistics are then computed. These include angular second

moment, contrast , correlation, inverse difference moment, entropy, sum entropy, difference entropy, and sum average. The statistics measure aspects of texture in the scene and allow the authors to represent various hair and skin textures of the human face. The statistics are appropriately scaled in order to produce scale invariance. This is combined with the mean and standard deviation of gray levels in each picture to form a ten dimensional input vector.

The authors employ cascade correlation and Kohonen-like networks to test these primitives. They report that both networks are able to correctly determine the presence of a face in an image 77% to 83% of the time. This accuracy could be increased with the use of higher order statistics in the input vector. Due to the presentation of faces for training from a variety of vantage points, the primitive form appears to uphold rotation, scale, and translation invariance. The basic disadvantage to this approach is that it is computationally expensive and extremely time consuming.

Terry Caelli, Mario Ferraro, and Erhardt Barth present a hierarchically organized set of explicit primitive representations in [3]. The representation is based on the relationships of basic parts of a pattern. Unary elements are defined as the basic constructs of any pattern. Binary features are used to define the relationships of these unary predicates. The degree of invariance upheld by

the primitives is dependent on the choice of unary and binary elements. The authors next extend these primitives for use in a three dimension object recognition application [3]. The scheme is labeled a natural representation due to its biological roots. Objects in an image are defined by a set of two dimensional structures. These structures are in turn defined by sets of tangent planes, surface normals, and rates of normal change (curvature). The planes are physically represented in sets of differential operators. This is consistent with how the human visual system is believed to organize its primitives. An assumption is made that low-level processes exist to convert light intensity and depth information into a three dimensional coordinate system. It is possible that this is also accomplished in the human visual system in the vertical columns of the primary visual cortex. The authors claim that the representation can uphold scale, rotation, and translation invariance.

4.3 Algorithm Structure

This next section will examine two particular algorithm organizations utilized in object detection. The goal of an algorithm structure should be to adequately separate discriminating features of the particular data representation

employed. The structures in this section employ a variation of the Kohonen feature map discussed in chapter 3. The variation includes a hierarchy of feature maps similar to what appears below in Figure 4.1 for a collection of 1 dimensional maps.

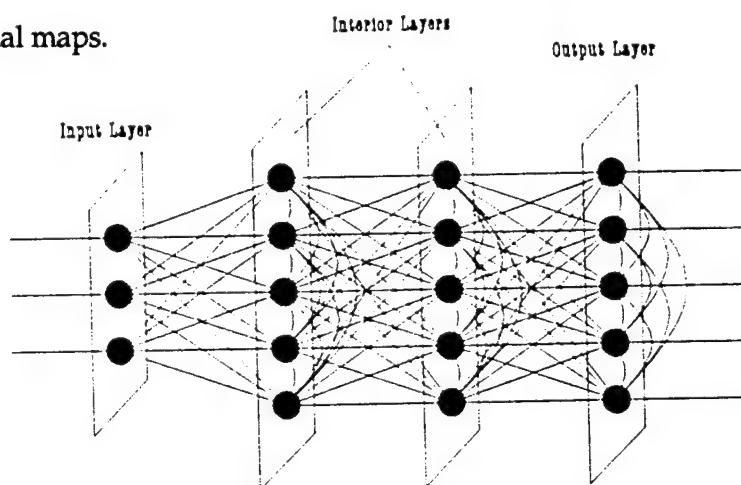


Figure 4.1 The Hierarchical Kohonen Feature Map

Ralf Linsker first explored the benefits of multilevel feature maps [18]. Benefits stem from the greater number of processing levels and the system's enhanced ability to organize along more complicated feature qualities. Each layer is able to represent progressively more complex features. One of the strengths that Linsker points out is that each level in the hierarchy preserves the information content of the input data. Each level only effects the information organization. Linsker also finds some surprising similarities between these networks and their biological counterparts. He found that lower levels of a multilevel network were

able to self-organize into activation regions similar to center-surround receptive fields found in the human visual system. Higher layers in the network produced outputs that were orientation-selective. This is the same selectivity that exists in the brain's cortical columns. Linsker's research defines the fundamental strength of unsupervised networks. The basic difficulty in organizing features is determining what information is important in producing a desired output for the network. Which connections need to be adjusted and how? This is the classic credit assignment problem. Unsupervised networks avoid the issue by preserving all of the information locally at each level. This reinforces the need for a fully descriptive primitive representation. The primitive captures all of the information of the image that will ever be used in the network. The grouping of local information allows for feature discrimination at the global level where similar features are located in neighboring neurons in the network. The grouping of information in the network does allow for discrimination between features. Linsker refers to this overall goal of a network as variance maximization. The quality of the representation scheme determines whether or not discriminating between these features has the same meaning as discriminating between objects in the real world.

There are two biologically inspired models that have evolved over the last few years that exploit the qualities highlighted by Linsker. These models build on various aspects of the human visual system. James Landowski and Baldamar Gil write that biologically motivated systems like Adaptive Resonance Theory (ART) and Neocognitron are extremely useful for applications in modern perception systems [17]. ART tends to be too complicated for practical implementation. Additionally, it is not shift invariant and is not easily applied for use in parallel implementations. This makes its use somewhat expensive and time consuming. ART has been successfully applied in [4], [5], and [29].

Landowski and Gill apply a discrete Neocognitron model based on the human retina for use in a target recognition system [17]. The system is intended to recognize the presence of tanks on a battlefield. The model consists of multiple layers of self-organizing feature maps. Infrared signatures are the inputs to the system which employs Hebbian learning in its self-organization. Training is accomplished level by level beginning with the input layer. Conventional training techniques recommend that an input pattern be presented to the first layer and that a single learning cycle be performed. The same pattern is then presented to an additional layer as a second learning cycle is performed. This continues until all of the levels have been trained on the

pattern. The authors report that this technique often requires up to sixty epochs for the network to reach its saturation state. An alternative method is proposed in which ten to twenty epochs are performed on each layer prior to training on additional layers. The result is a more robust system that has organized in a smaller amount of time.

The outputs of this system are complex discriminating features but they are not able to recognize targets [17]. The authors point out that the system is not able to incorporate relative feature importance based on apriori information. It is also not able to accumulate evidence in the identification of a tank's presence. Finally, the system is not able to support a small representation of target information. An additional mechanism is incorporated into the overall system that can include this type of information in the identification process. The authors use a fuzzy logic system to perform this task. In this way, the strength of the self-organized network in isolating important features can still be used. Fuzzy logic uses probability distribution functions to incorporate apriori and circumstantial information. This is in fact an example of the optimal application of an unsupervised neural network. It adds further evidence to Erkki Oja's comment that neural networks are best suited for intermediate levels of vision processing. Higher levels of processing like association require state-

space searching or symbolic processing [21]. This is not efficiently implemented with a neural network. These are, however, the specific tasks that AI or fuzzy logic systems have been designed for. The authors agree with this finding and do construct a robust system that is capable of achieving positive recognition rates of close to 100%. The system is even able to achieve this level of performance with mobile targets.

Hayder Ali Alkasimi introduces another form of the multilevel Neocognitron model for object detection applications [1]. He calls this a dynamic Neocognitron model. The network is designed to improve the practicality of using self-organizing networks for recognizing multiple objects. Current computer architectures are faced with an $O(n^2)$ requirement in growth for each additional object that is memorized by the network. The model consists of three layers of feature maps. The first two layers are connected like standard Kohonen feature maps. The first layer is the input plane. The next layer is the edge plane. Alkasimi specifies the output of this layer to consist of standardized edges and thus prevents the employment of local Hebbian learning. An edge is defined to be two neighboring pixels oriented in one of four standard directions (vertical, horizontal, and both diagonals). This sets the edge layer size at four times that of the input plane.

The dynamic portion of the system is the output layer. During training patterns to be recognized are presented to the system. This produces a certain number of activated neurons in the edge plane. The network then allocates a neuron in the output layer. The author next defines a feature explicitly as a combination of two or more edges. The particular inter-relationships of all activated edges for the training pattern are then stored in a local list at the allocated neuron. These are the stored features that can be used for discrimination in recognition tasks. The author claims that the stored connectivity of the edges allows for inherent rotation and translation invariance in the system. The connectivity is simply a representation of an object's occluding contour. Unfortunately, occluding contours can be extremely misleading for 3 dimensional objects. The network is not able to support scale invariance since different sizes of the same contour will result in different features. The prevention of Hebbian learning (and subsequent definition of simple primitives representations) prevents information preservation at each level. Organization is specified for the first two levels. The system assumes that all pertinent data will reside in only these types of features. The system is, therefore, limited to basic pattern recognition tasks. During actual operation, output layer neurons denote the degree of presence for each stored object. This

application demonstrates the importance of Hebbian learning in allowing a network to self-organize along optimal feature differences.

4.4 Dedicated Hardware

Current research into the field of object recognition has produced at least one piece of dedicated hardware. G. Sandini and M. Tistarelli report on the construction of a space-variant sensor device designed and patterned after the retina of the human eye [28]. The device picks up high resolution information at the center of its fovea and lower resolution information on the outskirts. The authors point out that high resolution data often results in information saturation of a recognition system. In most cases, the information is totally unrelated to the particular object of interest. The retina in the human eye allows for selective areas of high resolution. Information is gathered in areas where it is most likely to be relevant. The implication of such a device for artificial systems is that the recognition process becomes significantly less computationally expensive. A large portion of unrelated data is filtered out of the system prior to processing. The central problem for the device is how to select important areas of a scene. This remains a significant area of frustration

for researchers. Currently gaze control is limited to the directed movements of an observer.

Chapter 5

Conclusion

This report has given an overview of the implementation of current object detection schemes and some of the more important considerations involved. It has shown how an understanding of human vision can lend greater insight into the construction of such systems. Several imaging principles were addressed in an attempt to better isolate and define the perception problem. Invariance was seen as a key factor for reliable recognition. The methods of perception employed in biology served as a starting point for studying artificial systems. This led to exploration in the field of self-organized neural networks. These networks were seen to organize data in a similar fashion as the human visual system based on local relationships. Kohonen feature maps were examined in order to identify some basic properties of self-organization that are useful in recognition tasks. Finally, current applications in terms of primitive representation and algorithm organization were addressed. These applications showed that neural networks do not entirely solve the perception problem. The

strength and value of an unsupervised neural network remains its ability to discriminate between various features in the input space. The lesson to be learned here is that neural networks are extremely useful for certain types of data processing. Current technology is not able to imitate the human visual system. Research into neural technology is valuable for two reasons. First, a system may be built using some aspect of this technology. Second, research into neural networks may give added insight into better ways of approaching the perception problem. In any case, neural networks are valuable in developing any future technology that may be able to mimic the human visual system.

Bibliography

- [1] Hayder Ali Alkasimi. *Explorations in Cognitive Processing and Visual Neural Recognition*. Masters thesis, The University of Texas at Austin, Department of Electrical and Computer Engineering, December 1991.
- [2] Marijke F. Augusteijn and Tammy L. Skufca. "Identification of Human Faces through Texture-Based Feature Recognition and Neural Network Technology." In 1993 *IEEE International Conference on Neural Networks*, vol. 1, pp. 392-398, 1993.
- [3] T. Caelli, M. Ferraro, and E. Barth. "Aspects of Invariant Pattern and Object Recognition." In Harry Wechsler, editor, *Neural Networks for Perception; Volume 1; Human and Machine Perception*, pp. 234-247. New York: Academic Press, Inc, 1992.
- [4] Gail A. Carpenter, Stephan Grossberg, and John Reynolds. "A Neural Network Architecture for Fast On-line Supervised Learning and Pattern Recognition." In Harry Wechsler, editor, *Neural Networks for Perception; Volume 1; Human and Machine Perception*, pp. 248-264. New York: Academic Press, Inc, 1992.
- [5] Gail A. Carpenter. "Neural Network Models for Pattern Recognition and Associative Memory." In Gail A. Carpenter and Stephen Grossberg, editors, *Pattern Recognition by Self-Organizing Neural Networks*, pp. 1-34. Cambridge: The MIT Press, 1991.
- [6] Jinhui Chao, Kenji Minowa, and Shigeo Tsujii. "Unsupervised Learning of 3D Objects Conserving Global Topological Order." In 1992 *IEEE International Conference on Systems Engineering*, pp. 24-27, 1992.
- [7] Leon N. Cooper. "Visual Cortex: Window on the Biological Basis of Learning and Memory." In Harry Wechsler, editor, *Neural Networks for Perception; Volume 1; Human and Machine Perception*, pp. 8-24. New York: Academic Press, Inc, 1992.
- [8] Francis Crick and Christof Koch. "The Problem of Consciousness." In *Scientific American*, pp. 153-159, September 1992.

- [9] S. Edelman. "A Network model of Object Recognition in Human Vision." In Harry Wechsler, editor, *Neural Networks for Perception; Volume 1; Human and Machine Perception*, pp. 8-24. New York: Academic Press, Inc, 1992.
- [10] Arthur P. Ginsburg and William R. Hendee. "Quantification of Visual Capability." In William R. Hendee and Peter N.T. Wells, editors, *The Perception of Visual Information*, pp. 52-72. New York: Springer-Verlag, 1993.
- [11] Stephen Grossberg. "Neural Pattern Discrimination." In Gail A. Carpenter and Stephen Grossberg, editors, *Pattern Recognition by Self-Organizing Neural Networks*, pp. 1-34. Cambridge: The MIT Press, 1991.
- [12] E. Harth, K.P. Unnikrishnan, and A.S. Pandya. "The Inversion of Sensory Processing by Feedback Pathways: A Model of Visual Cognitive Functions." In *Science*, vol. 237, pp. 184-187, July 1987.
- [13] Simon Haykin. *Neural Networks; A Comprehensive Foundation*. New York: Macmillian College Publishing Company, 1994.
- [14] Robert Hecht-Nielsen. *Neurocomputing*. New York: Addison-Wesley Publishing Company, 1991.
- [15] David H. Hubel and Torsten N. Wiesel. "Brain Mechanisms of Vision; A functional architecture that may underlie processing of sensory information in the cortex is revealed by studies of the activity and the organization in space of neurons in the primary visual cortex." In *Scientific American*, vol. 241, no. 3, pp. 150-163, 1983.
- [16] Teuvo Kohonen. "Self-organized Formation of Topologically Correct Feature Maps." In *Biological Cybernetics*, vol. 43, pp. 59-69.
- [17] James G. Landowski and Baldamar Gil. "Application of a Vision Neural Network in an Automatic Target Recognition System." In Steven K. Rogers, editor, *SPIE-The International Society for Optical Engineering*, vol. 1709, part 1, pp. 34-43, 1992.

- [18] Ralph Linsker. "Self-organization in a Perceptual Network." In *IEEE Computer*, vol. 21, no. 3, pp. 105-117, 1988.
- [19] D.M. MacKay. "Ways of Looking at Perception." In Weiant Wathen-Dunn, editor, *Models for the Perception of Speech and Visual Form*, pp. 25-43. Cambridge: The M.I.T. Press, 1967.
- [20] David Marr. *Vision; A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman and Company, 1982.
- [21] Erkki Oja. "Self-Organizing Maps and Computer Vision." In Harry Wechsler, editor, *Neural Networks for Perception; Volume 1; Human and Machine Perception*, pp. 368-385. New York: Academic Press, Inc, 1992.
- [22] D.I. Perrett, M.H. Harries, R. Bevan, S. Thomas, P.J. Benson, A.J. Mistlin, A.J. Chitty, J.K. Hietanen, and J.E. Ortega. "Frameworks of Analysis for the Neural Representations of Animate Objects and Actions." In *Journal of Experimental Biology*, vol. 146, pp. 87-113, 1989.
- [23] D.I. Perrett, A.J. Mistlin, and A.J. Chitty. "Visual Neurons Responsive to Faces." In *Trends in Neuroscience*, vol. 10, no. 9, pp. 358-363, 1987.
- [24] D.I. Perrett, P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, and M.A. Jeeves. "Visual Cells in the Temporal Cortex Sensitive to Face View and Gaze Direction." In *Proceedings of the Royal Society of London*, vol. 223, pp. 293-317, 1985.
- [25] Ronald R. Price. "Image Manipulation." In William R. Hendee and Peter N.T. Wells, editors, *The Perception of Visual Information*, pp. 202-229. New York: Springer-Verlag, 1993.
- [26] Helge Ritter, Thomas Martinetz, and Klaus Schulten. *Neural Computation and Self-Organizing Maps*. New York: Addison-Wesley Publishing Company, 1992.

- [27] Bart M. Ter Haar Romeny and Luc Florack. "A Multiscale Geometric Model of Human Vision." In William R. Hendee and Peter N.T. Wells, editors, *The Perception of Visual Information*, pp. 73-114. New York: Springer-Verlag, 1993.
- [28] G. Sandini and M. Tistarelli. "Vision and Space-Variant Sensing" In Harry Wechsler, editor, *Neural Networks for Perception; Volume 1; Human and Machine Perception*, pp. 398-425. New York: Academic Press, Inc, 1992.
- [29] Michael Seibert and Allen M. Waxman. "Learning and Recognizing 3D Objects from Multiple Views in a Neural System." In Harry Wechsler, editor, *Neural Networks for Perception; Volume 1; Human and Machine Perception*, pp. 426-444. New York: Academic Press, Inc, 1992.
- [30] Peter F. Sharp and Russel Philips. "Physiological Optics." In William R. Hendee and Peter N.T. Wells, editors, *The Perception of Visual Information*, pp. 1-51. New York: Springer-Verlag, 1993.
- [31] Janusz A. Starzyk and Sinkuo Chai. "Vector Contour Representation for Object Recognition in Neural Networks." In 1992 IEEE International Conference on Systems, Man and Cybernetics, vol. 1, pp. 399-403, 1992.
- [32] Leonard Uhr. *Pattern Recognition, Learning, and Thought; Computer-Programmed Models of Higher Mental Processes*. Englewood Cliffs: Prentice-Hall, Inc., 1973.
- [33] Roger J. Watt. *Understanding Vision*. New York: Academic Press, 1991.
- [34] Harry Wechsler. "Multiscale and Distributed Visual Representations and Mappings for Invariant Low-Level Perception." In Harry Wechsler, editor, *Neural Networks for Perception; Volume 1; Human and Machine Perception*, pp. 462-476. New York: Academic Press, Inc, 1992.

Vita

Steacy Housholder was born in New Brighton, Pennsylvania on October 30th, 1971, the son of Steacy and Karen Housholder. He completed his secondary education in 1990 at Blue Valley North High School in Overland Park, Kansas. He next attended the United States Air Force Academy in Colorado Springs, Colorado. In 1994, he graduated with honors and received a Bachelor of Science in Electrical Engineering with a minor in German. As a commissioned officer, he entered graduate school at the University of Texas at Austin in August of 1994.

Permanent Address: 12637 Grandview
Overland Park, Kansas
66213